OPT2025: 17th Annual Workshop on Optimization for Machine Learning

On the Limits of Momentum in Decentralized and Federated Optimization

Riccardo Zaccone RICCARDO.ZACCONE@POLITO.IT
Politecnico di Torino

Sai Praneeth Karimireddy KARIMIRE@USC.EDU
USC Viterbi School of Engineering

Carlo Masone CARLO.MASONE@POLITO.IT
Politecnico di Torino

Abstract

Recent works have explored the use of momentum in local methods to enhance distributed SGD. This is particularly appealing in Federated Learning (FL), where momentum intuitively appears as a solution to mitigate the effects of statistical heterogeneity. Despite recent progress in this direction, it is still unclear if momentum can guarantee convergence under unbounded heterogeneity in decentralized scenarios, where only some workers participate at each round. In this work we analyze momentum under cyclic client participation, and theoretically prove that it remains inevitably affected by statistical heterogeneity. Similarly to SGD, we prove that decreasing step-sizes do not help either: in fact, any schedule decreasing faster than $\Theta\left(1/t\right)$ leads to convergence to a constant value that depends on the initialization and the heterogeneity bound. Numerical results corroborate the theory, and deep learning experiments confirm its relevance for realistic settings.

1. Introduction

Modern deep learning applications demand intensive training on large amount of data, often distributed across decentralized silos or user personal devices. To address such system constraints and comply with data regulations, learning algorithms have evolved towards more advanced and flexible systems that enable decentralized training at a global scale. In such systems, not all workers participate at each training step, due to local faults, network issues or simply temporary unavailability. Moreover, they cannot usually exchange their data, either because of efficiency or privacy concerns. These are the main premises of Federated Learning (FL), a paradigm focused on privacy-preserving training from decentralized data. Algorithms of this kind usually consist of an iterative two-step process involving 1) local training at client-side, each on its own private data, and 2) global optimization at the server, using aggregated local updates. While this scheme promotes efficiency by looser synchronization, *statistical heterogeneity* among clients' data and *partial client participation* expose the optimization to *client drift* and biased server updates.

Aiming for an effective solution to these problems, research has recently shifted towards extending momentum [18] to distributed algorithms. For example, a plethora of momentum-based FL algorithms have been proposed to overcome the adverse effects of data heterogeneity [1, 5, 10, 14, 17, 19, 23, 24]. Similarly, momentum is appealing in distributed learning to reduce the overall com-

munication overhead [21], and recently has been scaled up to more decentralized environments [2]. However, on a theoretical level, we only have a partial understanding of how momentum affects convergence in a decentralized regimen. [1] proved that momentum can converge under unbounded heterogeneity when all clients participate at each round (*full participation*). [24] went a step further, proposing a novel Generalized Heavy-Ball Momentum (GHBM) formulation that achieves the same convergence guarantees but with a more general *cyclic partial participation* assumption. Yet, it is unclear whether the same result can be further extended to classical momentum under the same cyclic partial participation assumption and without bounded heterogeneity.

This work provides a clear answer to this question: *can (classical) momentum enable convergence under unbounded heterogeneity in decentralized settings with partial participation?* The answer is negative: even with (classical) momentum, the convergence rate relies on the heterogeneity bound. This further confirms that GHBM [24] is, to the best of our knowledge, the only momentum-based distributed algorithm circumventing this limitation. Related works are deferred to Appendix A.1.

2. The Effect of Heterogeneity on Momentum

We study the effect of momentum in heterogeneous settings by considering a minimal setup with two heterogeneous clients. Our analysis is based on modeling the algorithm dynamics as a discrete-time linear system, and it reveals a clear decomposition: the *zero-input response* captures objectives shared by all clients, while the *zero-state response* isolates heterogeneous ones. This formulation unveils the source of convergence limitations and the role of heterogeneity in the system's behavior.

2.1. Preliminaries

Notation. We use $T \in \mathbb{N}$ to denote total number of iterations of the algorithms, with $[T]$ representing the set $\{1, 2, ..., T\}$ and $t \in [T]$ the t-th iteration. We denote as $f(\theta)$ the objective function parametrized by model parameters $\theta \in \mathbb{R}^d$, where d is the dimensionality of the model. We indicate with $\mathcal{S}$ the set of all clients and with $\mathcal{S}^t \subset \mathcal{S}$ the ones active at t-th iteration. Throughout the paper, to express the asymptotic growth rate of the convergence rates, we use $\mathcal{O}, \Theta$ and Ω to respectively indicate an upper, exact bound and a lower bound, with symbols hiding constant factors.

Setting. We consider a distributed learning system where a set $\mathcal{S}$ of clients collaboratively solve a learning problem. This can be formalized as a finite-sum optimization problem, where an objective function $f(\theta)$ is expressed in terms of function components $f_i(\theta)$, with each client optimizing a different component. Formally, the objective of the algorithm is finding:

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^d} \left[f(\theta) := \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} f_i(\theta) \right] \tag{1}$$

Gradient-Based Methods with Momentum. In modern deep learning applications, permutation-based variants of gradient descent (GD) are the most common algorithms. They reduce the computational burden by sampling and calculating a gradient over a function component f_i at each step, mainly differing by the strategy used to select the component. Among those, Stochastic Gradient Descent (SGD) and Incremental Gradient Descent (IGD) are most popular: SGD samples f_i uniformly and randomly, while IGD fixes any permutation of function components and samples cyclically from it. In this context, momentum has been used as a mechanism to reduce the impact of *noise* introduced by sampling, and improve convergence. Momentum consists in a moving average of past gradients, and it is often regarded as a way to reduce the variance of model updates [13].

Formally, the update rule of GD variants with momentum in its *heavy-ball* form can be written as:

$$m^t \leftarrow (\theta^{t-1} - \theta^{t-2}), \qquad \theta^t \leftarrow \theta^{t-1} - \eta(1-\beta)\nabla f^t(\theta^{t-1}) + \beta m^t \tag{2}$$

where η is the step-size, $\beta \in [0, 1)$ is the momentum factor and $f^t(\theta)$ is the component at time t.

From Centralized to Decentralized Algorithms. In the context of decentralized and federated learning, clients often represent function components. This analogy is rooted in the fact that data among clients are expected to differ. At each round $t \in [T]$, a fraction of $C \in (0, 1]$ clients $\mathcal{S}^t$ is selected for training. These clients may take gradients over mini-batch of data or additionally run an optimization algorithm locally over multiple local steps, and send back aggregated updates. In this work we assume clients that are sampled cyclically and take just one step of GD. Mini-batches and local steps improve the computational and communication efficiency but they may (i) introduce additional noise at the client side and (ii) cause forms of *client-drift*, especially under heterogeneity.

2.2. Assumptions

We assume objective functions are μ-strongly convex, with clients sampled in a fixed cyclic order (Thms. 1 and 3). Heterogeneity is captured by a bound on gradient dissimilarity between local and global objectives (Thm. 2), and we study how the convergence rate depends on it.

Assumption 1 (Strong Convexity) *Let it be a constant $\mu > 0$, then for any i, θ_1, θ_2 the following holds:*

$$f_i(\theta_2) \geq f_i(\theta_1) + \langle \nabla f_i(\theta_1), \theta_2 - \theta_1 \rangle + \frac{\mu}{2} \|\theta_2 - \theta_1\|^2$$

Assumption 2 (Bounded Gradient Dissimilarity) *There exist a constant $G \geq 0$ such that, $\forall i$, θ:*

$$\frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla f_i(\theta) - \nabla f(\theta)\| \leq G$$

Assumption 3 (Cyclic Participation) *Let $\mathcal{S}^t$ be the set of clients sampled at any round t. A sampling strategy is "cyclic" with period $p = 1/C$ if:*

$$\mathcal{S}^t \equiv \mathcal{S}^{t-p} \qquad \forall\, t > p \quad \wedge \quad \mathcal{S}^k \cap \mathcal{S}^t = \varnothing \qquad \forall\, k \in (t-p, t)$$

2.3. Learning Problem Construction

The intuition suggesting the use of momentum in a decentralized setting is that, being a moving average of past gradients, momentum achieves variance reduction effects [13]. We construct a learning problem which should be favorable to momentum under partial client participation. The global objective function is composed by only two objectives selected cyclically, one each round.

Lemma 4 (IGD with momentum on two one-dimensional clients) *For any positive constants G, μ, define μ-strongly convex functions $f_1(\theta) := \frac{\mu}{2}\theta^2 + G\theta$ and $f_2(\theta) := \frac{\mu}{2}\theta^2 - G\theta$ satisfying assumption 2 and such that $f(\theta) = \frac{1}{2}(f_1(\theta) + f_2(\theta))$. Under cyclic participation (assumption 3) with $C = 0.5$, for any $t \geq 1$ the evolution of* IGD *with momentum, with step-size η_t and momentum weight β is described by a discrete-time linear system with state-space representation:*

$$\begin{cases} \mathbf{z}[t] = \Psi(t,1)\mathbf{z}[1] + \sum_{k=2}^{t} \Psi(t,k)\mathbf{B}\mathbf{u}[k] \\ \mathbf{y}[t] = \mathbf{C}\Psi(t,1)\mathbf{z}[1] + \mathbf{C}\sum_{k=2}^{t} \Psi(t,k)\mathbf{B}\mathbf{u}[k] \end{cases}$$

where, given $\tilde{\eta}_t = (1-\beta)\eta_t$:

$$\mathbf{z}[t] = \begin{pmatrix} \theta^t & \theta^{t-1} \end{pmatrix}^\top, \qquad \mathbf{u}[t] = \left((-1)^t \tilde{\eta}_t G \right), \qquad \mathbf{A}[t] = \begin{pmatrix} 1+\beta-\mu\tilde{\eta}_t & -\beta \\ 1 & 0 \end{pmatrix}$$

$$\mathbf{B} = \begin{pmatrix} 1 & 0 \end{pmatrix}^\top, \qquad \mathbf{C} = \begin{pmatrix} 1 & 0 \end{pmatrix}, \qquad \Psi(t,k) := \textstyle\prod_{s=k+1}^{t} \mathbf{A}[s]$$

Both the minimum number of objective function components and the cyclic sampling are supposed to represent the easiest scenario for momentum, since we can guarantee that we observe the global objective every two rounds, ensuring momentum does not get biased towards either of the components. Lem. 4 shows that our learning problem can be analyzed with a discrete-time linear system, where $\mathbf{z}[t]$, $\mathbf{u}[t]$ and $\mathbf{y}[t] = \theta^t$ are respectively the state, the input and the output at t-th round. Let us express $f_{1,2}(\theta) = f_{hom}(\theta) + f_{het}(\theta)$, where $f_{hom}(\theta) = f(\theta)$ and $f_{het}(\theta) = \pm G\theta$. Then, the update w.r.t. the shared objective maps onto the *natural response* of the system, while ∇f_{het} appears as external force of the system, acting as "disturb signal" to the optimization of the global objective. This offers immediate understanding of the impact of noise on the algorithm's convergence, whether it is stochastic as in SGD or deterministic as in IGD. If there was no input (*i.e.* $f_{1,2}(\theta)$ were homogeneous or both sampled at each round), then convergence would depend only on initial conditions, with an exponentially fast rate under proper constant step-size $\eta_t = \eta$. Conversely, the presence of heterogeneity leads to a convergence rate determined by how the terms related to the initial conditions (called *zero-input response*) and the input (called *zero-state response*) interact: this depends on the choice of step-size η_t, which enters in the state matrix and as scaling to the input.

2.4. Convergence under Constant Step-sizes

The following theorem reveals that, similarly as it is known for vanilla SGD and IGD, the addition of momentum does not bring any asymptotic advantage in the lower bound of the convergence rate. This directly implies that algorithms based on classical momentum *cannot* be used to provide strong theoretical guarantees against statistical heterogeneity in decentralized and federated learning settings under partial participation (*i.e.* an heterogeneity bound G is still necessary). Since our analysis does not take into account local steps, the following theorem holds for both FEDAVGM and FEDCM under cyclic client participation.

Theorem 5 *For any positive constants G, μ there exist μ-strongly convex functions satisfying assumption 2 for which, under proper constant step-size η and for any momentum factor $\beta \in [0,1)$, the output of* FEDCM *and* FEDAVGM *under cyclic partial participation (assumption 3), has the following asymptotic error:*

$$f(\theta^t) - f(\theta^*) = \Theta\left(\frac{G^2}{\mu T^2}\right)$$

The proof is deferred to Appendix C.2.

The result is based on the analysis of the LTI system resulting from Lem. 4 and the assumption of proper constant step-size $\eta_t = \eta$. Under these conditions, the zero-input response converges exponentially fast, the faster the higher the step-size, matching the known convergence rate of GD. Conversely, the zero-state response converges to a 2-period cycle limit, whose amplitude is proportional to the step-size. As a consequence, the final asymptotic rate is dominated by the response to the input, and a step-size as small as $\eta = \mathcal{O}(1/T)$ must be imposed to obtain a linear rate.

2.5. Convergence under Decreasing Step-sizes

The intuitive reason for adopting decreasing step-sizes lies on the observation that heterogeneity enters the optimization as an external input, scaled by the effective step-size $\eta_t(1-\beta)$. This suggest that decreasing η_t over time may offer a benefit not visible when it is kept constant. We study the problem in Lem. 4 under a polynomial decreasing step-size schedule of the type $\eta_t = \eta/t^\alpha$, where t is the current iteration and $\alpha > 0$ is an hyperparameter controlling the decay rate of η_t. The following theorem reveals that, even when η_t is decreasing, the dependence on the heterogeneity bound cannot be eliminated, and that overly fast-decaying step-size schedules are detrimental.

Theorem 6 *For any positive constants G, μ, α there exist μ-strongly convex functions satisfying assumption 2 for which, under decreasing step-size $\eta_t \sim \mathcal{O}(1/t^\alpha)$,the output of* FEDCM *and* FEDAVGM *under cyclic participation (assumption 3), even assuming initialization at optimum ($\theta^0 = \theta^*$), has the following error:*

$$f(\theta^t) - f(\theta^*) = \begin{cases} \Theta\left(\dfrac{G^2}{\mu t^{2\alpha}}\right) & \text{if } 0 < \alpha < 1 \\ \Theta\left(\dfrac{G^2}{\mu t^{2\min(\mu\eta,1)}}\right) & \text{if } \alpha = 1 \\ \Theta\left(\dfrac{G^2}{\mu}\right) & \text{if } \alpha > 1 \end{cases}$$

The proof is deferred to Appendix C.2.

Slowly-decreasing step-sizes. When the decay rate of the step-size is sufficiently slow (*i.e.* $0 < \alpha < 1$), the convergence rate is strictly slower than in Thm. 5, as $2\alpha < 2$, and the dependence on the heterogeneity bound G remains. From the mathematical point of view, the bottleneck in the rate arises from the solution of the zero-input response, which decays as a polynomial in α, while the zero-state response still decays exponentially fast. As such, for large t the rate is dominated by the former term, and the final convergence value θ^t is the same irrespective of initial conditions θ^0.

Fast-decreasing step-sizes. When $\alpha = 1$, the convergence rate depends on the choice of initial step size η. When a small $\eta < 1/\mu$ is chosen, the rate depends on $\mu\eta$, getting slower as η is chosen smaller. On the other hand, when a large $\eta \geq 1/\mu$ is chosen, the rate matches the one in Thm. 5. Similar findings have been observed for SGD under the same step-size schedule by [9]. Mathematically, the transition between $t^{-\mu\eta}$ to t^{-1} in the rate arises because the state transition matrix $\Psi(t,s)$ now decays only polynomially to zero, not exponentially as in the previous case. As is, the rate now depends on how the zero-input and zero-state responses interact: when $\eta < 1/\mu$, a term depending on the initialization affects the rate, so θ^t will depend on θ^0. On the contrary, when $\eta > 1/\mu$, the rate is dominated only by the response to heterogeneity.

Overly fast-decreasing step-sizes. When the step-size decays faster than linearly, the algorithm fails to reach an arbitrarily small optimality gap. Both the solutions of the homogeneous and heterogeneous part of the system in Lem. 4 are affected, because the state transition matrix does not longer decay to zero. This means that, not only the zero-state response converges to a constant depending on G, but also the the zero-input response converges to a constant depending on the initialization.

3. Numerical Results

We provide numerical results confirm our theoretical findings, evaluating three step-size schedules: constant (as analyzed in Thm. 5), polynomially decreasing ($\eta_t = \eta/t^\alpha$, with $\alpha > 0$, as in Thm. 6), and exponentially decreasing ($\eta_t = \eta\gamma^t$, with $\gamma \in (0,1)$). The experiments, shown in Tab. 1

Table 1: **Effect function heterogeneity and decreasing step-size on IGD with (left) and without momentum (right):** θ^t after $T = 10^6$ iterations for the problem in Lem. 4. Heterogeneity affects convergence linearly, and step-size schedules decaying faster than $\Theta\left(1/t^\alpha\right)$ lead to worse solutions, both when not starting at the optimum (*i.e.* $\theta^0 \neq 0$ and $G = 0$) and when objectives are heterogeneous (*i.e.* $G > 0$ and $\theta^0 = \theta^*$).

STEP-SIZE SCHEDULE	$G = 100$		$G = 10$		$G = 0$	$G = 100$		$G = 10$		$G = 0$
	$\theta^0 = 0$	$\theta^0 = 10$	$\theta^0 = 0$	$\theta^0 = 10$	$\theta^0 = 10$	$\theta^0 = 0$	$\theta^0 = 10$	$\theta^0 = 0$	$\theta^0 = 10$	$\theta^0 = 10$
CONSTANT	2.5e−05	2.5e−05	2.5e−06	2.5e−06	5.7e−08	1.5e−05	3.7e+00	1.5e−06	3.7e+00	3.7e+00
POLYNOMIAL										
$\alpha = 0.1$	7.2e+00	7.2e+00	7.2e−01	7.2e−01	−5.0e−324	7.2e+00	7.2e+00	7.2e−01	7.0e−01	5.0e−324
$\alpha = 0.5$	2.5e−02	2.5e−02	2.5e−03	2.5e−03	1.5e−323	2.5e−02	2.5e−02	2.5e−03	2.5e−03	9.4e−322
$\alpha = 1$	2.5e−05	2.5e−05	2.5e−06	2.5e−06	−1.7e−78	−8.6e−06	−8.3e−06	−8.6e−07	−6.0e−07	2.6e−07
$\alpha = 2$	4.8e+01	5.7e+01	4.8e+00	1.4e+01	9.0e+00	−1.9e+01	−1.9e+01	−1.9e+00	−1.8e+00	1.0e−01
EXPONENTIAL										
$\gamma = 0.9999$	1.5e−17	1.5e−17	2.4e−18	2.4e−18	−1.5e−323	1.9e−17	1.9e−17	2.3e−18	2.3e−18	0.0e+00
$\gamma = 0.999$	1.8e−17	1.8e−17	−6.3e−18	−6.3e−18	−1.7e−163	2.3e−17	2.3e−17	7.0e−18	7.0e−18	0.0e+00
$\gamma = 0.99$	1.1e−14	9.8e−15	1.1e−15	−3.5e−16	−1.5e−15	1.5e−16	1.5e−16	1.8e−18	1.8e−18	1.1e−65
$\gamma = 0.9$	−7.2e+00	−6.2e+00	−7.2e−01	−2.4e+01	9.5e−01	−1.0e−04	−8.2e−05	−1.0e−05	1.1e−05	2.2e−05

for comparison shown for both with/without momentum, confirm that momentum is affected by heterogeneity, and that that fast-decaying schedules negatively affect convergence to the optimum. Experiments on a non-convex realistic FL setting are deferred to Appendix A.3

Constant and Slowly-decreasing Step-sizes. Results in Tab. 1 show that, when the learning rate is constant or slowly decreasing (*i.e.* $\alpha < 1$), the final value at convergence always linearly depends on the heterogeneity bound G, and it is irrespective of initialization. This validates the theory, which predicts an exponential decay rate of the initial conditions and a linear decay of the perturbation caused by heterogeneity. The result of constant learning rate and linear decay ($\alpha = 1$) are equal in all cases but when the system is homogeneous (*i.e.* $G = 0$): in this case, since the decay rate of the initial conditions is exponential, a bigger step-size is better This motivates why the smaller the decay, the closer the solution is to the optimum, which is contrary to the heterogeneous cases.

Fast-decreasing Step-sizes. When $\alpha = 1$, the decay rate of initialization and heterogeneity interact, as they are both polynomial, and the overall rate depends on the choice of the step-size. Indeed, as shown in Tab. 2, when $\eta > 1/\mu$ the solution depends on the heterogeneity, since the decay rate of the initialization is $\mathcal{O}(t^{-\mu\eta})$, which is faster than $\mathcal{O}(t^{-1})$: this makes the solution independent of θ^0 On the contrary, $\eta < 1/\mu$ the decay rate of the initialization is slower than $\mathcal{O}(t^{-1})$, so the final solution θ^t is different for $\theta^0 = 0$ and $\theta^0 = 10$. When the step-size decay rate is too fast, the system does not converge to the optimum, but to a final value depending on initialization and heterogeneity, as highlighted by the red rows in Tab. 1.

Table 2: **Impact of step-size with fast decaying polynomial schedule on IGD with momentum:** θ^t after $T = 10^6$ iterations for the learning problem in Lem. 4, with $\epsilon = 10^{-2}$.

POLYNOMIAL DECAY RATE	INITIAL STEP-SIZE	$G = 10$	
		$\theta^0 = 0$	$\theta^0 = 10$
$\alpha = 1$	$\eta = \frac{1(1+\beta)}{\mu(1-\beta)} - \epsilon$	2.5e−06	2.5e−06
$\alpha = 1$	$\eta = \frac{1}{\mu} - \epsilon$	−3.9e−06	−1.2e−04

4. Conclusions

This paper addresses a gap in understanding the role of momentum in distributed optimization with statistical heterogeneity and partial worker participation. While momentum is appealing to build robustness to statistical heterogeneity, our work demonstrates that it does not inherently overcome the challenges posed by heterogeneous data. By unveiling this fundamental limitation, this work provides a more realistic basis for its use in heterogeneous decentralized environments.

Funding

Riccardo Zaccone and Carlo Masone declare that financial support was received for the research, authorship, and/or publication of this article. This study was carried out within the project FAIR - Future Artificial Intelligence Research - and received funding from the European Union Next-GenerationEU [PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013 - CUP: E13C22001800001]. This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

[1] Ziheng Cheng, Xinmeng Huang, Pengfei Wu, and Kun Yuan. Momentum benefits non-iid federated learning simply and provably. In *ICLR*, 2024.

[2] Arthur Douillard, Qixuan Feng, Andrei Alex Rusu, Rachita Chhaparia, Yani Donchev, Adhiguna Kuncoro, MarcAurelio Ranzato, Arthur Szlam, and Jiajun Shen. Diloco: Distributed low-communication training of language models. In *2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ICML 2024)*, 2024.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXv:1512.03385*, 2015.

[4] Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-IID data quagmire of decentralized machine learning. In *ICML*, 2020.

[5] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

[6] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Federated visual classification with real-world data distribution. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, 2020.

[7] Yerlan Idelbayev. Proper ResNet implementation for CIFAR10/CIFAR100 in PyTorch, 2021.

[8] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[9] Arnulf Jentzen and Philippe von Wurstemberger. Lower error bounds for the stochastic gradient descent optimization algorithm: Sharp convergence rates for slowly and fast decaying learning rates. *Journal of Complexity*, 2020.

[10] Geeho Kim, Jinkyu Kim, and Bohyung Han. Communication-efficient federated learning with accelerated client gradient. In *CVPR*, 2024.

[11] Yujun Kim, Jaeyoung Cha, and Chulhee Yun. Incremental gradient descent with small epoch counts is surprisingly slow on ill-conditioned problems. In *ICML*, 2025.

[12] Anastasia Koloskova, Nikita Doikov, Sebastian U Stich, and Martin Jaggi. On convergence of incremental gradient for non-convex smooth functions. In *ICMLR*, 2024.

[13] Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. In *NeurIPS*, 2020.

[14] Yixing Liu, Yan Sun, Zhengtao Ding, Li Shen, Bo Liu, and Dacheng Tao. Enhance local consistency in federated learning: A multi-step inertial momentum approach, 2023. URL https://arxiv.org/abs/2302.05726.

[15] Zijian Liu and Zhengyuan Zhou. On the last-iterate convergence of shuffling gradient methods. In *ICML*, 2024.

[16] PHUONG_HA Nguyen, Lam Nguyen, and Marten van Dijk. Tight dimension independent lower bound on the expected convergence rate for diminishing step sizes in sgd. In *NeurIPS*, 2019.

[17] Emre Ozfatura, Kerem Ozfatura, and Deniz Gündüz. Fedadc: Accelerated federated learning with drift control. In *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021.

[18] Boris Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr Computational Mathematics and Mathematical Physics*, 1964.

[19] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. In *ICLR*, 2021.

[20] Itay Safran and Ohad Shamir. How good is sgd with random shuffling? In *Proceedings of Thirty Third Conference on Learning Theory*, 2020.

[21] Jianyu Wang, Vinayak Tantia, Nicolas Ballas, and Michael Rabbat. Slowmo: Improving communication-efficient distributed sgd with slow momentum. In *ICLR*, 2020.

[22] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018.

[23] Jing Xu, Sen Wang, Liwei Wang, and Andrew Chi-Chih Yao. Fedcm: Federated learning with client-level momentum. *arXiv preprint arXiv:2106.10874*, 2021.

[24] Riccardo Zaccone, Sai Praneeth Karimireddy, Carlo Masone, and Marco Ciccone. Communication-efficient heterogeneous federated learning with generalized heavy-ball momentum. *Transactions on Machine Learning Research*, 2025.

Appendix A. Additional Discussion

A.1. Related Works

Gradient Descent (GD) and its variants have long been objective of study in the context of finite-sum optimization problems. Restricting the gradient calculation to single function components (*i.e.* a small subset of data) at each iteration, those methods trade off noisy updates for computational efficiency. Most of the analyses address SGD or shuffling gradient methods [12, 15, 20]. [9] provides sharp lower bounds on SGD for decreasing step-sizes, while [16] prove dimension-independent lower bounds over all possible sequences of diminishing step-sizes. The recent work of [11] studies the convergence rate of IGD at small iteration count.

While in all cases an heterogeneity bound is necessary, the above works consider algorithms *without* momentum. Since it has been proved that momentum has a variance reduction effect [13], it is not clear i) if the fundamental reliance on the heterogeneity remains even with momentum, and ii) if decreasing step-sizes play a role. In this work we analyze the simplest setting in which momentum could intuitively bring an advantage w.r.t. heterogeneous objectives: as we show, this corresponds to an instance of the IGD algorithm *with* momentum.

A.2. Circumventing the Momentum Lower Bounds

The findings in this section confirm classical momentum cannot be employed in decentralized learning to completely overcome the effects of statistical heterogeneity. To the best of authors' knowledge, the only momentum-based algorithm circumventing this limitation is the Generalized Heavy-Ball Momentum (GHBM) [24]. As authors explain, leveraging an incremental aggregated gradient perspective, its momentum update rule approximates the one classical momentum has in full participation. Therefore, the limitations we refer to in this paper do not apply to GHBM.

A.3. Federated Learning Experiments.

Following the experimental protocol of [24], we present results on a realistic FL setting, comparing vanilla FEDAVG to FEDCM, a FL algorithm based on classical momentum, under constant step-size. As shown in Fig. 1, classical momentum demonstrates ineffective in high heterogeneous decentralized settings with partial participation, failing at improving FEDAVG. This further confirms the relevance of our findings for realistic scenarios.

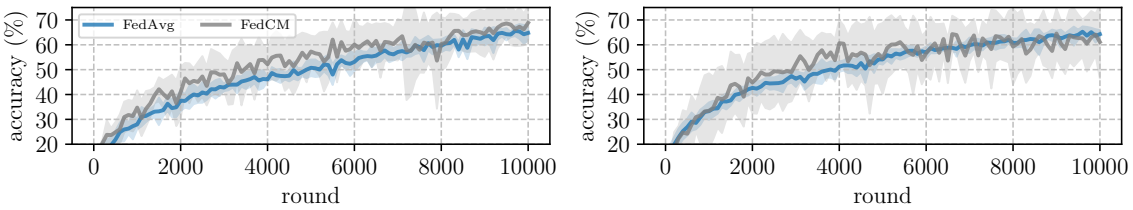

Figure 1: **FEDAVG and FEDCM under cyclic participation:** under high heterogeneity and partial participation, FL-methods based on classical momentum do not offer a substantial improvement over simpler methods without momentum. Results on CIFAR-10 with RESNET-20 (left) and CNN (right). The reference accuracy in centralized settings is $\approx 86\%$ for CNN and $\approx 89\%$ for RESNET-20.

Appendix B. Experimental Setting

Datasets and Models. We consider CIFAR-10 to experiment with image classification tasks, each one respectively having 10 and 100 classes. For all methods, training images are preprocessed by applying random crops, followed by random horizontal flips. Both training and test images are finally normalized according to their mean and standard deviation. As the main model for experimentation, we used a model similar to LENET-5 as proposed in [6]. To further validate our findings, we also employed a RESNET-20 as described in [3], following the implementation provided in [7]. Since batch normalization [8] layers have been shown to hamper performance in learning from decentralized data with skewed label distribution [4], we replaced them with group normalization [22], using two groups in each layer.

Hyperparameters. As per the hyperparameters, for FEDAVG and CNN we search the server step-size $\eta \in \{2, 1.5, 1, 0.5, 0.1\}$ and local step-size $\eta_l \in \{0.1, 0.05, 0.01, 0.005\}$ and found the best performing to be $\eta = 1.5$ and $\eta_l = 0.01$. For RESNET-20, we search the server step-size $\eta \in \{1.5, 1, 0.1\}$ and local step-size $\eta_l \in \{1, 0.5, 0.1, 0.01\}$ and found the best performing to be $\eta = 1$ and $\eta_l = 0.5$. Similarly, for FEDCM and CNN we search the server step-size $\eta \in \{1, 0.5, 0.1, 0.05\}$ and local step-size $\eta_l \in \{1, 0.5, 0.1, 0.05\}$ and found the best performing to be $\eta = 0.1$ and $\eta_l = 0.1$. For RESNET-20, we search the server step-size $\eta \in \{1.5, 1, 0.5, 0.1\}$ and local step-size $\eta_l \in \{1, 0.5, 0.1, 0.5\}$ and found the best performing to be $\eta = 1$ and $\eta_l = 0.1$. The momentum factor is searched among $\beta \in \{0.95, 0.9, 0.85\}$ and set as $\beta = 0.9$.

Simulating Heterogeneity. We simulate arbitrary heterogeneity by splitting the total datasets according to a Dirichlet distribution with concentration parameter α, following [6]. In practice, we draw a multinomial $q_i \sim \mathbf{Dir}(\alpha p)$ from a Dirichlet distribution, where p describes a prior class distribution over N classes, and α controls the heterogeneity among all clients: the greater α the more homogeneous the clients' data distributions will be. After drawing the class distributions q_i, for every client i, we sample training examples for each class according to q_i without replacement.

The experiments provided in the main paper adopt $\alpha = 0$, following the experimental setting of [24].

Metrics and Experimental protocol. We consider the model accuracy in predicting the correct class images belong to. Results are always reported as average of 5 independent runs, with stardard deviation directly shown in Fig. 1.

Compute resource. Deep learning experiments in Fig. 1 require one hour of training each with a commodity GPU (in our case, a NVIDIA GTX1070). Theoretical experiments run on CPU in some seconds each.

Appendix C. Deferred Proofs

C.1. Auxiliary Lemmas

Here is a collection of some smaller technical lemmas that are used within the proofs of the main results.

Lemma 7 *Let $f(x)$ be a non-negative, monotonically decreasing function that is integrable over an interval $[a, b]$, where $a < b$ are integers. The following inequality holds:*

$$\sum_{k=a+1}^{b} f(k) \leq \int_{a}^{b} f(x)\, dx$$

Proof Since $f(x)$ is a monotonically decreasing function on the interval $[a, b]$, for any integer $k \in [a+1, b]$, and for any $x \in [k-1, k]$, we have that:

$$f(k) \leq f(x) \Rightarrow \int_{k-1}^{k} f(k)\, dx \leq \int_{k-1}^{k} f(x)\, dx \tag{3}$$

Since $f(k)$ is constant w.r.t. the integration variable x, we have that:

$$f(k) \leq \int_{k-1}^{k} f(x)\, dx \tag{4}$$

Summing up from $k = a+1$ to $k = b$ and using the additive property of integrals:

$$\sum_{k=a+1}^{b} f(k) \leq \sum_{k=a+1}^{b} \int_{k-1}^{k} f(x)\, dx \tag{5}$$

$$= \int_{a}^{b} f(x)\, dx \tag{6}$$

■

Lemma 8 *Let $\eta < \frac{2^{\alpha}}{\mu}$, and let the function $\Psi_1(t, s, \alpha)$ be:*

$$\Psi_1(t, s, \alpha) := \prod_{k=s+1}^{t} \left(1 - \frac{\mu\eta}{k^{\alpha}}\right)$$

Then, for any $t >= 2$ *and* $1 \leq s < t$ *the following holds:*

$$\Psi_1(t,s,\alpha) \leq \begin{cases} \exp\left(-\mu\eta \dfrac{t^{1-\alpha}-s^{1-\alpha}}{1-\alpha}\right) & \text{if } 0<\alpha<1 \\ \left(\dfrac{s}{t}\right)^{\mu\eta} & \text{if } \alpha = 1 \\ \exp\left(-\dfrac{\mu\eta}{2^\alpha}\right) & \text{if } \alpha > 1 \end{cases}$$

$$\Psi_1(t,s,\alpha) \geq \begin{cases} \exp\left(-\dfrac{2^\alpha\mu\eta}{2^\alpha-\mu\eta}\dfrac{t^{1-\alpha}-s^{1-\alpha}}{1-\alpha}\right) & \text{if } 0<\alpha<1 \\ \left(\dfrac{s}{t}\right)^{\frac{2\mu\eta}{2-\mu\eta}} & \text{if } \alpha = 1 \\ \exp\left(-\dfrac{2^\alpha\mu\eta}{2^\alpha-\mu\eta}\zeta_H(\alpha,s+1)\right) & \text{if } \alpha > 1 \end{cases}$$

where $\zeta_H(\alpha,s) := \sum_{k=s}^{\infty} \frac{1}{k^\alpha}$ *is the Hurwitz zeta-function.*

Proof **Case** $0 < \alpha \leq 1$: For the upper bound, we have that

$$\begin{aligned} \Psi_1(t,s,\alpha) &= \prod_{k=s+1}^{t}\left(1-\frac{\mu\eta}{k^\alpha}\right) && (7)\\ &= \exp\left(\sum_{k=s+1}^{t}\ln\left(1-\frac{\mu\eta}{k^\alpha}\right)\right) && (8)\\ &\leq \exp\left(-\sum_{k=s+1}^{t}\frac{\mu\eta}{k^\alpha}\right) && (9)\\ &\overset{7}{\leq} \exp\left(-\mu\eta\int_s^t \frac{1}{k^\alpha}\,dk\right) && (10)\\ &= \begin{cases} \exp\left(-\mu\eta\dfrac{t^{1-\alpha}-s^{1-\alpha}}{1-\alpha}\right) & \text{if } 0<\alpha<1 \\ \exp\left(-\mu\eta\log\left(\dfrac{t}{s}\right)\right) = \left(\dfrac{s}{t}\right)^{\mu\eta} & \text{if } \alpha = 1 \end{cases} && (11) \end{aligned}$$

where in the step (9) we used the inequality $\ln(1-x) \leq -x$ for $x > 0$, with $x = \frac{\mu\eta}{k^\alpha}$. Similarly, for the lower bound we have that

$$\Psi_1(t,s,\alpha) = \prod_{k=s+1}^{t} \left(1 - \frac{\mu\eta}{k^\alpha}\right) \tag{12}$$

$$= \exp\left(\sum_{k=s+1}^{t} \ln\left(1 - \frac{\mu\eta}{k^\alpha}\right)\right) \tag{13}$$

$$\geq \exp\left(\sum_{k=s+1}^{t} \underbrace{-\frac{1}{1-\mu\eta/k^\alpha}}_{\text{increasing}} \frac{\mu\eta}{k^\alpha}\right) \tag{14}$$

$$\overset{k\geq s+1\geq 2}{\geq} \exp\left(-\frac{2^\alpha}{2^\alpha - \mu\eta} \sum_{k=s+1}^{t} \frac{\mu\eta}{k^\alpha}\right) \tag{15}$$

$$\overset{7}{\geq} \exp\left(-\frac{2^\alpha \mu\eta}{(2^\alpha - \mu\eta)} \int_s^t \frac{1}{k^\alpha}\, dk\right) \tag{16}$$

$$= \begin{cases} \exp\left(-\dfrac{2^\alpha\mu\eta}{2^\alpha - \mu\eta} \dfrac{t^{1-\alpha} - s^{1-\alpha}}{(1-\alpha)}\right) & \text{if } 0 < \alpha < 1 \\ \exp\left(-\dfrac{2\mu\eta}{2-\mu\eta} \ln\left(\dfrac{t}{s}\right)\right) = \left(\dfrac{s}{t}\right)^{\frac{2\mu\eta}{2-\mu\eta}} & \text{if } \alpha = 1 \end{cases} \tag{17}$$

where in the step (14) we used the inequality $\log(1-x) \geq -\frac{x}{1-x}$ for $x > 0$, with $x = \frac{\mu\eta}{k^\alpha}$.

Case $\alpha > 1$: For $\alpha > 1$, we have that

$$(s+1)^{-\alpha} \leq \sum_{k=s+1}^{t} k^{-\alpha} < \sum_{k=s+1}^{\infty} k^{-\alpha} < \infty \;\forall\, t \tag{18}$$

Therefore, for the upper bound we have

$$\Psi_1(t,s,\alpha) \overset{\text{Eq. (9)}}{\leq} \exp\left(-\sum_{k=s+1}^{t} \frac{\mu\eta}{k^\alpha}\right) \overset{\text{Eq. (18)}}{\leq} \exp\left(-\frac{\mu\eta}{(s+1)^\alpha}\right) \overset{s\geq 1}{\leq} \exp\left(-\frac{\mu\eta}{2^\alpha}\right) \tag{19}$$

For the lower bound we have

$$\Psi_1(t,s,\alpha) \overset{\text{Eq. (15)}}{\geq} \exp\left(-\frac{2^\alpha\mu\eta}{2^\alpha - \mu\eta} \sum_{k=s+1}^{t} \frac{1}{k^\alpha}\right) \tag{20}$$

$$\overset{\text{Eq. (18)}}{\geq} \exp\left(-\frac{2^\alpha\mu\eta}{2^\alpha - \mu\eta} \sum_{k=s+1}^{\infty} \frac{1}{k^\alpha}\right) \tag{21}$$

$$= \exp\left(-\frac{2^\alpha\mu\eta}{2^\alpha - \mu\eta} \zeta_H(\alpha, s+1)\right) \tag{22}$$

■

Corollary 9 *Let $\eta < \frac{2^\alpha}{\mu}$, and let the function $\Psi_1(t,s,\alpha)$ be:*

$$\Psi_1(t,s,\alpha) := \prod_{k=s+1}^{t} \left(1 - \frac{\mu\eta}{k^\alpha}\right)$$

Then the following holds:

$$\lim_{t\to\infty} \Psi_1(t,s,\alpha) = \lim_{t\to\infty} \begin{cases} \exp\left(-t^{1-\alpha}\right) & \text{if } 0 < \alpha < 1 \\ \left(\frac{1}{t}\right)^{\mu\eta} & \text{if } \alpha = 1 \end{cases}$$

Moreover, for $\alpha > 1$, it holds that:

$$\lim_{t\to\infty} \Psi_1(t,s,\alpha) = c, \qquad c \in \left(\exp\left(-\frac{2^\alpha\mu\eta}{2^\alpha - \mu\eta}\zeta_H(\alpha, s+1)\right), \exp\left(-\frac{\mu\eta}{2^\alpha}\right)\right)$$

Proof The proof of the statement follows from taking the limit for $t \to \infty$, for each range of α, of upper bounds in Lem. 8, which give the slowest decay. ■

Lemma 10 *Let μ, η and β positive constants, and let the function $\Psi_2(t,s,\alpha)$ be:*

$$\Psi_2(t,s,\alpha) := \prod_{k=s+1}^{t} \left(\beta + \frac{\mu\eta\beta}{k^\alpha}\right)$$

For any $1 \le s < t$ and $\alpha > 0$, the following holds:

$$\Psi_2(t,s,\alpha) \le \begin{cases} \beta^{t-s} \exp\left(\mu\eta\frac{t^{1-\alpha} - s^{1-\alpha}}{1-\alpha}\right) & \text{if } 0 < \alpha < 1 \\ \beta^{t-s}\left(\frac{t}{s}\right)^{\mu\eta} & \text{if } \alpha = 1 \\ \beta^{t-s}\exp\left(\mu\eta\zeta_H(\alpha, s+1)\right) & \text{if } \alpha > 1 \end{cases}$$

$$\Psi_2(t,s,\alpha) > 0$$

where $\zeta_H(\alpha, s) := \sum_{k=s}^{\infty} \frac{1}{k^\alpha}$ is the Hurwitz zeta-function.

Proof The fact that $\Psi_2(t,s,\alpha)$ is positive (lower bound) is trivial. For the upper bound, we can write

$$\Psi_2(t,s,\alpha) = \prod_{k=s+1}^{t} \left(\beta + \frac{\mu\eta\beta}{k^\alpha}\right) \tag{23}$$

$$= \beta^{t-s} \prod_{k=s+1}^{t} \left(1 + \frac{\mu\eta}{k^\alpha}\right) \tag{24}$$

$$= \beta^{t-s} \exp\left(\sum_{k=s+1}^{t} \ln\left(1 + \frac{\mu\eta}{k^\alpha}\right)\right) \tag{25}$$

$$\le \beta^{t-s} \exp\left(\sum_{k=s+1}^{t} \frac{\mu\eta}{k^\alpha}\right) \tag{26}$$

where in the last step we used the inequality $\ln(x) \leq x - 1 \ \forall x > 0$, with $x = \left(1 + \frac{\mu\eta}{k^\alpha}\right) > 0$, which is always verified since $\mu, \eta, k > 0$. Now, we differentiate the next steps depending on the value of α.

Case $0 < \alpha < 1$: Since the function within the summation in Eq. (26) is decreasing, we use Lem. 7:

$$0 < \Psi_2(t, s, \alpha) \leq \beta^{t-s} \exp\left(\mu\eta \sum_{k=s+1}^{t} \frac{1}{k^\alpha}\right) \tag{27}$$

$$\overset{7}{\leq} \beta^{t-s} \exp\left(\mu\eta \int_s^t \frac{1}{k^\alpha}\, dk\right) \tag{28}$$

$$= \beta^{t-s} \exp\left(\mu\eta \frac{t^{1-\alpha} - s^{1-\alpha}}{1-\alpha}\right) \tag{29}$$

Case $\alpha = 1$: Using Lem. 7 as in the previous case, we have that:

$$0 < \Psi_2(t, s, \alpha) \leq \beta^{t-s} \exp\left(\mu\eta \sum_{k=s+1}^{t} \frac{1}{k^\alpha}\right) \tag{30}$$

$$\overset{7}{\leq} \beta^{t-s} \exp\left(\mu\eta \int_s^t \frac{1}{k}\, dk\right) \tag{31}$$

$$= \beta^{t-s} \exp\left(\mu\eta \ln\left(\frac{t}{s}\right)\right) \tag{32}$$

$$= \beta^{t-s} \left(\frac{t}{s}\right)^{\mu\eta} \tag{33}$$

Case $\alpha > 1$:

$$0 < \Psi_2(t, s, \alpha) \leq \beta^{t-s} \exp\left(\mu\eta \sum_{k=s+1}^{t} \underbrace{\frac{1}{k^\alpha}}_{>0}\right) \tag{34}$$

$$\leq \beta^{t-s} \exp\left(\mu\eta \sum_{k=s+1}^{\infty} \frac{1}{k^\alpha}\right) \tag{35}$$

$$= \beta^{t-s} \exp\left(\mu\eta\zeta_H(\alpha, s+1)\right) \tag{36}$$

This concludes the proof. ■

Corollary 11 *Let μ, η and β positive constants, and let the function $\Psi_2(t, s, \alpha)$ be:*

$$\Psi_2(t, s, \alpha) := \prod_{k=s+1}^{t} \left(\beta + \frac{\mu\eta\beta}{k^\alpha}\right)$$

Then the following holds:

$$\lim_{t\to\infty} \Psi_2(t, s, \alpha) = \lim_{t\to\infty} \beta^t$$

Proof The proof of the statement follows from taking the limit for $t \to \infty$, for each range of α, of both upper and lower bounds of $\Psi_2(t, s, \alpha)$ in Lem. 10 and using the squeeze theorem. ■

Lemma 12 *Let $\Psi_1(t, s, \alpha)$ as defined in Lem. 8, and let the summation $S(t, \alpha)$ be:*

$$S(t, \alpha) := \sum_{s=2}^{t} \Psi_1(t, s, \alpha) \frac{1}{s^{2\alpha}}$$

Then, for any $\alpha > 1$ the following holds:

$$\frac{\exp\left(-\frac{2^\alpha \mu\eta}{2^\alpha - \mu\eta}\zeta_R(\alpha)\right)}{2\alpha - 1} \leq \lim_{t\to\infty} S(t, \alpha) \leq \zeta_R(2\alpha)$$

Proof For $\alpha > 1$, we have that:

$$S(t, \alpha) \overset{8}{\leq} \sum_{s=2}^{t} \frac{1}{s^{2\alpha}} \exp\left(-\frac{\mu\eta}{2^\alpha}\right) \tag{37}$$

$$\leq \sum_{s=1}^{\infty} \frac{1}{s^{2\alpha}} = \zeta_R(2\alpha) \tag{38}$$

On the other hand:

$$S(t, \alpha) \overset{8}{\geq} \sum_{s=2}^{t} \frac{1}{s^{2\alpha}} \exp\left(-\frac{2^\alpha \mu\eta}{2^\alpha - \mu\eta}\zeta_H(\alpha, s)\right) \tag{39}$$

$$\geq \exp\left(-\frac{2^\alpha \mu\eta}{2^\alpha - \mu\eta}\zeta_R(\alpha)\right) \sum_{s=2}^{t} \frac{1}{s^{2\alpha}} \tag{40}$$

$$\geq \exp\left(-\frac{2^\alpha \mu\eta}{2^\alpha - \mu\eta}\zeta_R(\alpha)\right) \int_2^t \frac{1}{s^{2\alpha}}\, ds \tag{41}$$

$$= \exp\left(-\frac{2^\alpha \mu\eta}{2^\alpha - \mu\eta}\zeta_R(\alpha)\right) \frac{t^{1-2\alpha} - 2^{1-\alpha}}{1 - 2\alpha} \tag{42}$$

$$= \exp\left(-\frac{2^\alpha \mu\eta}{2^\alpha - \mu\eta}\zeta_R(\alpha)\right) \left(\frac{1}{2\alpha - 1} - \frac{2^{1-\alpha}}{(2\alpha - 1)t^{2\alpha-1}}\right) \tag{43}$$

Since $\lim_{t\to\infty} 1/t^{2\alpha-1} = 0$ because $2\alpha - 1 > 0$ since $\alpha > 1$. Putting together the results of Eq. (38) and (43), we have that:

$$\frac{\exp\left(-\frac{2^\alpha \mu\eta}{2^\alpha - \mu\eta}\zeta_R(\alpha)\right)}{2\alpha - 1} \leq \lim_{t\to\infty} S(t, \alpha) \leq \zeta_R(2\alpha) \tag{44}$$

■

Lemma 13 *Let $\alpha > 0, n > 0$ and $\Psi_2(t,s,\alpha) := \prod_{k=s+1}^{t}\left(\beta + \frac{\mu\eta\beta}{k^\alpha}\right)$. Then with $\beta \in [0,1)$ the following holds:*

$$\lim_{t\to\infty}\sum_{s=2}^{t}\Psi_2(t,s,\alpha)\frac{1}{s^n} = \lim_{t\to\infty}\frac{1}{t^n}$$

Proof For $\beta = 0$ the statement is trivially true, because $\Psi_2(t,s,\alpha) = 0$. Therefore, from this point on, we consider $\beta \in (0,1)$. For readability, let us define the shorthand notation for the quantity in the l.h.s. of the statement

$$S(t,\alpha,n) := \sum_{s=2}^{t}\Psi_2(t,s,\alpha)\frac{1}{s^n} \tag{45}$$

For any $\alpha > 0$ we have that:

$$\bar{S}(t,\alpha,n) := t^n S(t,\alpha) = \sum_{s=2}^{t}\underbrace{\prod_{k=s+1}^{t}\left(\beta + \frac{\mu\eta\beta}{k^\alpha}\right)\left(\frac{t}{s}\right)^n}_{>0} \tag{46}$$

To prove the statement we derive the convergence of $\bar{S}(t,\alpha,n)$, and consequently $S(t,\alpha,n)$, by using the **Dominated Convergence Theorem**. We recall that this theorem states that given a sequence $f_t(u)$ such that

cond. 1) $\lim_{t\to\infty} f_t(u) = f(u) < \infty$

cond. 2) there exist a summable function $g(u) \geq |f_t(u)|$

then it holds that:

$$\lim_{t\to\infty}\sum_{u=0}^{\infty} f_t(u) = \sum_{u=0}^{\infty}\lim_{t\to\infty} f_t(u) = \sum_{u=0}^{\infty} f(u) \tag{47}$$

We proceed to breakdown the analysis for different ranges of the variable α.

Case $0 < \alpha < 1$: Starting from Eq. (46), we have that:

$$\bar{S}(t,\alpha) = \sum_{s=2}^{t}\prod_{k=s+1}^{t}\left(\beta + \frac{\mu\eta\beta}{k^\alpha}\right)\left(\frac{t}{s}\right)^n \tag{48}$$

$$\overset{10}{\leq} \sum_{s=2}^{t}\beta^{t-s}\exp\left(\mu\eta\frac{t^{1-\alpha} - s^{1-\alpha}}{1-\alpha}\right)\left(\frac{t}{s}\right)^n \tag{49}$$

$$\overset{u:=t-s}{=} \sum_{u=0}^{t-2}\beta^{u}\exp\left(\mu\eta\frac{t^{1-\alpha} - (t-u)^{1-\alpha}}{1-\alpha}\right)\left(\frac{t}{t-u}\right)^n \tag{50}$$

$$\overset{c:=\frac{\mu\eta}{1-\alpha}}{=} \sum_{u=0}^{t-2}\underbrace{\beta^{u}\exp\left(c(t^{1-\alpha} - (t-u)^{1-\alpha})\right)\left(\frac{t}{t-u}\right)^n}_{:=f_t(u)} \tag{51}$$

For the *condition 1*, we have that

$$f(u) := \lim_{t\to\infty} \beta^u \exp\Big(c(t^{1-\alpha} - (t-u)^{1-\alpha})\Big)\left(\frac{t}{t-u}\right)^n = \beta^u$$

For the *condition 2*, we observe that:

$$f_t(u) = \underbrace{\beta^u}_{>0}\underbrace{\exp\left(c(t^{1-\alpha} - (t-u)^{1-\alpha})\right)}_{\geq 1}\underbrace{\left(\frac{t}{t-u}\right)^n}_{\geq 1}$$

and that the last two terms have a maximum in $u = t-2$. Thus, it follows that

$$\begin{aligned}\big|f_t(u)\big| &\leq \beta^u \exp\Big(c(t^{1-\alpha} - 2^{1-\alpha})\Big)\left(\frac{t}{2}\right)^n \\ &\leq \beta^u \exp\Big(ct^{1-\alpha}\Big)t^n := g(u)\end{aligned} \tag{52}$$

To verify that $g(u)$ is summable, we can apply the ratio test:

$$\lim_{u\to\infty}\left|\frac{g(u+1)}{g(u)}\right| = \lim_{u\to\infty}\frac{\beta^{u+1}\exp\Big(ct^{1-\alpha}\Big)t^n}{\beta^u \exp\Big(ct^{1-\alpha}\Big)t^n} = \beta$$

Since the ratio is $\beta < 1$, this confirms that $g(u)$ is summable. Therefore we obtain that:

$$\begin{aligned}\lim_{t\to\infty} S(t,\alpha,n) &= \lim_{t\to\infty} t^{-n}\sum_{u=0}^{\infty} f(u) \\ &= \lim_{t\to\infty}\frac{1}{(1-\beta)t^n} = \lim_{t\to\infty}\frac{1}{t^n}\end{aligned}$$

Case $\alpha = 1$ **:** For $\alpha = 1$ we proceed similarly to the previous case. From Eq. (46) we have that:

$$\bar{S}(t,1,n) = \sum_{s=2}^{t}\prod_{k=s+1}^{t}\left(\beta + \frac{\mu\eta\beta}{k^\alpha}\right)\left(\frac{t}{s}\right)^n \tag{53}$$

$$\overset{10}{\leq} \sum_{s=1}^{t}\beta^{t-s}\left(\frac{t}{s}\right)^{\mu\eta+n} \tag{54}$$

$$\overset{u:=t-s}{=} \sum_{u=0}^{t-2}\underbrace{\beta^u\left(\frac{t}{t-u}\right)^{\mu\eta+n}}_{:=f_t(u)\geq 1} \tag{55}$$

For the *condition 1*, we have that:

$$f(u) := \lim_{t\to\infty} f_t(u) = \beta^u$$

For the *condition 2*, the second term in $f_t(u)$ has a maximum in $u = t - 2$, i.e.,

$$\left|f_t(u)\right| \leq \beta^u \left(\frac{t}{2}\right)^{\mu\eta+n} \leq \beta^u t^{\mu\eta+n} := g(u)$$

Hence, going back to $S(t, \alpha)$ with Eq. (46) and (47), we have that:

$$\begin{aligned}
\lim_{t\to\infty} S(t,1,n) &= \lim_{t\to\infty} t^{-n}\bar{S}(t,1,n) \\
&= \lim_{t\to\infty} t^{-n} \sum_{u=0}^{\infty} f(u) \\
&= \lim_{t\to\infty} t^{-n} \sum_{u=0}^{\infty} \beta^u \\
&= \lim_{t\to\infty} \frac{1}{(1-\beta)t^n} = \lim_{t\to\infty} \frac{1}{t^n}
\end{aligned}$$

Case $\alpha > 1$ **:** The case $\alpha > 1$ is analogous, and differs from the above only for a constant factor. We have that:

$$\begin{aligned}
\bar{S}(t,\alpha,n) &= \sum_{s=2}^{t} \prod_{k=s+1}^{t} \left(\beta + \frac{\mu\eta\beta}{k^\alpha}\right) \left(\frac{t}{s}\right)^n \\
&\overset{10}{\leq} \sum_{s=2}^{t} \beta^{t-s} \exp\left(\mu\eta\zeta_H(\alpha, s+1)\right) \left(\frac{t}{s}\right)^n \\
&\overset{u:=t-s}{=} \sum_{u=0}^{t-2} \underbrace{\beta^u \exp\left(\mu\eta\zeta_H(\alpha, t-u+1)\right) \left(\frac{t}{t-u}\right)^n}_{:=f_t(u)}
\end{aligned}$$

For the *condition 1*, we have that $f(u) := \lim_{t\to\infty} f_t(u) = \beta^u$. For the *condition 2*, the maximum of the second and third terms of $f_t(u)$ is found at $u = t - 2$, hence we have

$$\left|f_t(u)\right| \leq \beta^u \exp\left(\mu\eta\zeta_H(\alpha, 3)\right) \left(\frac{t}{2}\right)^n \leq \beta^u \exp\left(\mu\eta\zeta_H(\alpha, 1)\right) t^n := g(u)$$

Finally, going back to $S(t, \alpha)$ with Eq. (46) and (47), we have that:

$$\begin{aligned}
\lim_{t\to\infty} S(t,\alpha,n) = \lim_{t\to\infty} t^{-n}\bar{S}(t,\alpha,n) &= \lim_{t\to\infty} t^{-n} \sum_{u=0}^{\infty} f(u) \\
&= \lim_{t\to\infty} t^{-n} \sum_{u=0}^{\infty} \beta^u \\
&= \lim_{t\to\infty} \frac{1}{(1-\beta)t^n} = \lim_{t\to\infty} \frac{1}{t^n}
\end{aligned}$$

■

Lemma 14 *Let $\alpha > 0, n > 0$ and $\Psi_2(t,s,\alpha) := \prod_{k=s+1}^{t}\left(\beta + \frac{\mu\eta\beta}{k^\alpha}\right)$. Then with $\beta \in [0,1)$ the following holds:*

$$\lim_{t\to\infty}\sum_{s=2}^{t}\Psi_2(t,s,\alpha)\frac{(-1)^s}{s^n} = \lim_{t\to\infty}\frac{(-1)^t}{t^n}$$

Proof For readability, let us define the shorthand notation for the quantity in the l.h.s. of the statement Notice that

$$S(t,\alpha,n) := \sum_{s=2}^{t}\Psi_2(t,s,\alpha)\frac{(-1)^s}{s^n} \tag{56}$$

and notice that:

$$|S(t,\alpha,n)| \leq \sum_{s=2}^{t}\Psi_2(t,s,\alpha)\frac{1}{s^n} \tag{57}$$

Therefore, from direct application of Lem. 13 on the r.h.s. and the squeeze-theorem we have that for any $\alpha > 0, n > 0$:

$$\lim_{t\to\infty} S(t,\alpha) = \pm\lim_{t\to\infty}\frac{1}{t^n} \tag{58}$$

$$= \lim_{t\to\infty}\frac{(-1)^t}{t^n} \tag{59}$$

■

Lemma 15 *Let $\alpha > 0$ and $\Psi_1(t,s,\alpha) := \prod_{k=s+1}^{t}\left(1 - \frac{\mu\eta}{k^\alpha}\right)$. Consider the function*

$$S(t,\alpha,n) := \sum_{s=2}^{t}\frac{(-1)^s}{s^n}\Psi_1(t,s,\alpha)$$

with $t \geq 2$ and $1 \leq s < t$. Then, for $n > 0$, the following holds:

$$\lim_{t\to\infty} S(t,\alpha,n) = \lim_{t\to\infty}\frac{(-1)^t}{t^n} \qquad \text{if } 0 < \alpha < 1$$

$$\lim_{t\to\infty} S(t,1,n) = \lim_{t\to\infty}\begin{cases}\dfrac{(-1)^t}{t^n} & \text{if } n < \mu\eta \\ \dfrac{1}{t^{\mu\eta}} & \text{otherwise}\end{cases} \qquad \text{if } \alpha = 1$$

$$\gamma_1 \exp\left(-\frac{2^\alpha\mu\eta}{2^\alpha - \mu\eta}\zeta_H(\alpha,3)\right) \leq \lim_{t\to\infty} S(t,\alpha,n) \leq \frac{1}{2^n}\exp\left(-\frac{\mu\eta}{2^\alpha}\right) \qquad \text{if } \alpha > 1$$

where $\gamma_1 := \left(\left(\frac{1}{2}\right)^n - \left(\frac{1}{3}\right)^n\frac{3^\alpha}{3^\alpha - \mu\eta}\right) > 0$ and $\zeta_H(\alpha,s) := \sum_{k=s}^{\infty}\frac{1}{k^\alpha}$ is the Hurwitz zeta-function.

Proof From the definition, rewrite $S(t,\alpha,n)$ as recurrence:

$$S(t,\alpha,n) = \sum_{s=2}^{t}\prod_{k=s+1}^{t}\left(1-\frac{\mu\eta}{k^\alpha}\right)\frac{(-1)^s}{s^n} \tag{60}$$

$$= \sum_{s=2}^{t-1}\prod_{k=s+1}^{t}\left(1-\frac{\mu\eta}{k^\alpha}\right)\frac{(-1)^s}{s^n} + \left[\frac{(-1)^s}{s^n}\prod_{k=s+1}^{t}\left(1-\frac{\mu\eta}{k^\alpha}\right)\right]_{s=t} \tag{61}$$

$$= \sum_{s=2}^{t-1}\prod_{k=s+1}^{t}\left(1-\frac{\mu\eta}{k^\alpha}\right)\frac{(-1)^s}{s^n} + \frac{(-1)^t}{t^n}\underbrace{\prod_{k=t+1}^{t}\left(1-\frac{\mu\eta}{k^\alpha}\right)}_{=1} \tag{62}$$

$$= \left(1-\frac{\mu\eta}{t^\alpha}\right)\sum_{s=2}^{t-1}\prod_{k=s+1}^{t-1}\left(1-\frac{\mu\eta}{k^\alpha}\right)\frac{(-1)^s}{s^n} + \frac{(-1)^t}{t^n} \tag{63}$$

$$= \left(1-\frac{\mu\eta}{t^\alpha}\right)S(t-1,\alpha,n) + \frac{(-1)^t}{t^n} \tag{64}$$

$$\overset{\text{unrolling}}{=} \prod_{k=3}^{t}\left(1-\frac{\mu\eta}{k^\alpha}\right)S(2,\alpha,n) + \sum_{k=3}^{t}\prod_{s=k+1}^{t}\left(1-\frac{\mu\eta}{s^\alpha}\right)\frac{(-1)^k}{k^n} \tag{65}$$

The solution of the above first-order non-homogeneous recurrence is the sum of the homogeneous solution $S^{(h)}(t,\alpha,n)$ and a particular solution $S^{(p)}(t,\alpha,n)$, which can be analyzed separately. From Eq. (65), we have that:

$$S^{(h)}(t,\alpha,n) = \prod_{k=3}^{t}\left(1-\frac{\mu\eta}{k^\alpha}\right)S(2,\alpha,n) \tag{66}$$

$$= \Psi_1(t,2,\alpha)\frac{1}{2^n} \tag{67}$$

For the particular solution, we look for a form $S^{(p)}(t,\alpha,n) = \gamma\frac{(-1)^t}{t^n}$, and by substituting into the original recurrence in Eq. (64) we have:

$$\gamma\frac{(-1)^t}{t^n} = \gamma\left(1-\frac{\mu\eta}{t^\alpha}\right)\frac{(-1)^{t-1}}{(t-1)^n} + \frac{(-1)^t}{t^n} \tag{68}$$

Dividing by $\frac{(-1)^t}{t^n}$:

$$\gamma = -\gamma\left(1-\frac{\mu\eta}{t^\alpha}\right)\left(\frac{t}{t-1}\right)^n + 1 \tag{69}$$

So, for $t\to\infty$, $\gamma\to\frac{1}{2}$ and:

$$\lim_{t\to\infty} S^{(p)}(t,\alpha,n) = \lim_{t\to\infty}\frac{1}{2}\frac{(-1)^t}{t^n} \tag{70}$$

The asymptotic behavior of the homogeneous solution, and so of the original recurrence, depends on α.

Case $0 < \alpha < 1$: From Eq. (67) and (70), we have that:

$$\lim_{t\to\infty} S(t,\alpha,n) \overset{9}{=} \lim_{t\to\infty} \left[\left(\frac{1}{2}\right)^n \exp\left(-t^{1-\alpha}\right) + \frac{1}{2}\frac{(-1)^t}{t^n}\right] \tag{71}$$

$$= \lim_{t\to\infty} \frac{(-1)^t}{t^n} \tag{72}$$

Case $\alpha = 1$: From Eq. (67) and (70), we have that:

$$\lim_{t\to\infty} S(t,\alpha,n) \overset{9}{=} \lim_{t\to\infty} \left[\left(\frac{1}{2}\right)^n \left(\frac{1}{t}\right)^{\mu\eta} + \frac{1}{2}\frac{(-1)^t}{t^n}\right] \tag{73}$$

$$= \lim_{t\to\infty} \begin{cases} \dfrac{(-1)^t}{t^n} & \text{if } n < \mu\eta \\ \dfrac{1}{t^{\mu\eta}} & \text{otherwise} \end{cases} \tag{74}$$

Case $\alpha > 1$: In this case the summation converges to a non-zero constant, and we use a different strategy. Starting from the original definition of $\Psi_1(t,s,\alpha)$ as per Lem. 8, we call $g(t,s) := \left(\frac{1}{s}\right)^n \prod_{k=s+1}^{t}\left(1-\frac{\mu\eta}{k^\alpha}\right)$. Noticing that the function is decreasing in s we have that:

$$S(t,\alpha,n) = \sum_{s=2}^{t} \frac{(-1)^s}{s^n} \prod_{k=s+1}^{t}\left(1-\frac{\mu\eta}{k^\alpha}\right) \tag{75}$$

$$= g(t,2) + \sum_{s=3}^{t} (-1)^s g(t,s) \tag{76}$$

$$\leq g(t,2) + \sum_{s=2}^{\lfloor t/2 \rfloor} \underbrace{\left(g(t,2s) - g(t,2s-1)\right)}_{<0} \tag{77}$$

$$\leq g(t,2) \overset{8}{\leq} \frac{1}{2^n} \exp\left(-\frac{\mu\eta}{2^\alpha}\right) \tag{78}$$

Similarly, from Eq. (75), we have that:

$$S(t,\alpha,n) = \sum_{s=2}^{t} \frac{(-1)^s}{s^n} \prod_{k=s+1}^{t}\left(1-\frac{\mu\eta}{k^\alpha}\right) \tag{79}$$

$$\geq \sum_{s=1}^{\lfloor t/2 \rfloor} \underbrace{\left(g(t,2s) - g(2s+1)\right)}_{>0} \tag{80}$$

$$\geq g(t,2) - g(t,3) \tag{81}$$

So, defining $\gamma_1 := \left(\left(\frac{1}{2}\right)^n - \left(\frac{1}{3}\right)^n \frac{3^\alpha}{3^\alpha - \mu\eta}\right) > 0$, we have that:

$$g(t,2) - g(t,3) = \left(\frac{1}{2}\right)^n \prod_{k=3}^{t}\left(1 - \frac{\mu\eta}{k^\alpha}\right) - \left(\frac{1}{3}\right)^n \prod_{k=4}^{t}\left(1 - \frac{\mu\eta}{k^\alpha}\right) \tag{82}$$

$$= \prod_{k=3}^{t}\left(1 - \frac{\mu\eta}{k^\alpha}\right)\left(\left(\frac{1}{2}\right)^n - \left(\frac{1}{3}\right)^n \left(1 - \frac{\mu\eta}{3^\alpha}\right)^{-1}\right) \tag{83}$$

$$= \prod_{k=3}^{t}\left(1 - \frac{\mu\eta}{k^\alpha}\right)\left(\left(\frac{1}{2}\right)^n - \left(\frac{1}{3}\right)^n \frac{3^\alpha}{3^\alpha - \mu\eta}\right) \tag{84}$$

$$\overset{8}{\geq} \gamma_1 \exp\left(\frac{2^\alpha \mu\eta}{2^\alpha - \mu\eta}\zeta_H(\alpha,3)\right) \tag{85}$$

■

C.2. Proofs of Main Theorems

Proof of Lem. 4 (IGD with momentum on two one-dimensional clients)

We assume each client is assigned one of the two below simple one-dimensional functions for any given μ and G, and assume functions are sampled cyclically, *i.e.*:

$$f^t(\theta) := \begin{cases} f_1(\theta) := \frac{\mu}{2}\theta^2 + G\theta & \text{if } t \text{ is odd} \\ f_2(\theta) := \frac{\mu}{2}\theta^2 - G\theta & \text{otherwise} \end{cases} \tag{86}$$

Both functions are μ-strongly convex and $f(\theta) = \frac{1}{2}\left(f_1(\theta) + f_2(\theta)\right) = \frac{\mu}{2}(\theta)^2$, which has global minimizer at $\theta^* = 0$. Running IGD with momentum, the update after each round is:

$$\theta^t \leftarrow \theta^{t-1} - \eta_t(1-\beta)\nabla f^t(\theta^{t-1}) + \beta\left(\theta^{t-1} - \theta^{t-2}\right) \tag{87}$$

$$= \theta^{t-1} - \underbrace{\eta_t(1-\beta)}_{:=\tilde{\eta}_t}\left(\mu\theta^{t-1} + (-1)^{t-1}G\right) + \beta\left(\theta^{t-1} - \theta^{t-2}\right) \tag{88}$$

$$= \theta^{t-1}\left(1 + \beta - \mu\tilde{\eta}_t\right) - \beta\theta^{t-2} + (-1)^t\tilde{\eta}_t G \tag{89}$$

We can formalize the analysis of the above as a second-order discrete-time linear system using state-space representation. A discrete-time linear system can be represented in state-space form as:

$$\begin{cases} \mathbf{z}[t] = \mathbf{A}[t]\mathbf{z}[t-1] + \mathbf{B}\mathbf{u}[t] \\ \mathbf{y}[t] = \mathbf{C}\mathbf{z}[t] \end{cases} \tag{90}$$

where:

$$\mathbf{z}[t] = \begin{pmatrix} z_1[t] & z_2[t] \end{pmatrix}^\top = \begin{pmatrix} \theta^t & \theta^{t-1} \end{pmatrix}^\top, \qquad \mathbf{u}[t] = (-1)^t\tilde{\eta}_t G$$

$$\mathbf{A}[t] = \begin{pmatrix} 1 + \beta - \mu\tilde{\eta}_t & -\beta \\ 1 & 0 \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} 1 & 0 \end{pmatrix}^\top, \qquad \mathbf{C} = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

Given an initial state condition $\mathbf{z}[1] = \begin{pmatrix} \theta^1 & \theta^0 \end{pmatrix}^\top$, with $\theta^1 = \theta^0$, the result of the lemma follows from unrolling the recursion and defining the state transition matrix $\Psi(t,k) := \prod_{s=k+1}^{t}\mathbf{A}[s]$.

Proof of Thm. 5 (Lower Bound under Constant Step-size)
Let $\mathbf{z}[t]$ and $\mathbf{y}[t]$ be the state-space representation and the output of the discrete linear time-invariant (LTI) system constructed in Eq. (90) of Lem. 4. We denote $\mathbf{y}_{ZIR}[t] := \mathbf{C}\Psi(t,1)\mathbf{z}[1]$ as the **zero-input response** and $\mathbf{y}_{ZSR}[t] := \mathbf{C}\sum_{k=1}^{t}\Psi(t,k)\mathbf{B}\mathbf{u}[k]$ as the **zero-state response**, which can be studied separately thanks to linearity.

We assume a constant step size, i.e., $\eta_t = \eta$ (or equivalently $\tilde{\eta}_t = \tilde{\eta}$) $\forall t$.

Solution of zero-input response. Under constant learning rate the state matrix $\mathbf{A}[t]$ is $\mathbf{A}[t] = \mathbf{A}\,\forall t$. Therefore the state-transition matrix becomes $\Psi(t,1) = \mathbf{A}^{t-1}$ and we have that $\mathbf{A}^{t-1}\mathbf{z}[1] \to 0 \iff \mathbf{A}^{t-1} \to 0$ as $t \to \infty$ for any given initial state $\mathbf{z}[1] \neq \mathbf{0}$. The asymptotic convergence of the response depends on the eigenvalues of the matrix $\mathbf{A}$ being strictly less than one. The eigenvalues of $\mathbf{A}$ are the solutions $\lambda_{1,2}$ to the associated characteristic equation, and to find the values of η, β which satisfy the condition we apply the Jury stability criterion:

$$P(\lambda) := \det(\lambda\mathbf{I} - \mathbf{A}) = \det\begin{pmatrix}\lambda - (1+\beta-\mu\tilde{\eta}) & \beta \\ -1 & \lambda\end{pmatrix} \tag{91}$$

$$= \lambda\left(\lambda - (1+\beta-\mu\tilde{\eta})\right) + \beta \tag{92}$$

$$= \lambda^2 - (1+\beta-\mu\tilde{\eta})\lambda + \beta \tag{93}$$

- *condition 1:* $\beta < |\mathbf{1}| : \beta < 1 \Rightarrow \beta \in [0,1)$ (94)
- *condition 2:* $\mathbf{P}(\mathbf{1}) > \mathbf{0}$:

$$\Rightarrow 1 - (1+\beta-\mu\tilde{\eta}) + \beta > 0 \quad \Rightarrow \mu(1-\beta)\eta > 0 \quad \Rightarrow \eta > 0 \tag{95}$$

- *condition 3:* $\mathbf{P}(-\mathbf{1}) > \mathbf{0}$:

$$\Rightarrow 1 + (1+\beta-\mu\tilde{\eta}) + \beta > 0 \quad \Rightarrow 2(1+\beta) > \mu\tilde{\eta} \quad \Rightarrow \eta < \frac{2(1+\beta)}{\mu(1-\beta)} \tag{96}$$

In the above steps we have used the definition $\tilde{\eta} := (1-\beta)\eta$ from Lem. 4. Summarizing, under the condition

$$\eta \in \left(0, \frac{2(1+\beta)}{\mu(1-\beta)}\right) \quad \text{with } \beta \in [0,1) \tag{97}$$

the norm of $\mathbf{y}_{ZIR}[t]$ is monotonically decreasing w.r.t. t and converges to zero as $t \to \infty$.

Solution of the zero-state response. Proceeding with the analysis of the zero-state response $\mathbf{y}_{ZSR}[t]$, we show that the presence of the periodic term (due to the cyclic client switching) induces an oscillatory dynamic that does not decrease to zero and that depends on G. Since the input is 2-periodic, the zero-state response converges to a limit cycle of the same period. Namely, for a some fixed $\mathbf{c} \in \mathbb{R}^2$, we search for a solution of the **periodic form** $\mathbf{z}[t] = (-1)^t\mathbf{c}$:

$$\mathbf{z}[t] = \mathbf{A}\mathbf{z}[t-1] + \mathbf{B}\mathbf{u}[t] \tag{98}$$

$$\overset{\text{periodic form}}{\Rightarrow} (-1)^t\mathbf{c} = \mathbf{A}(-1)^{t-1}\mathbf{c} + (-1)^t\tilde{\eta}G\mathbf{B} \tag{99}$$

$$\overset{\text{division by } (-1)^t}{\Rightarrow} \mathbf{c} = -\mathbf{A}\mathbf{c} + \tilde{\eta}G\mathbf{B} \tag{100}$$

$$\overset{\text{group } \mathbf{c} \text{ to l.h.s}}{\Rightarrow} \mathbf{c} = (\mathbf{I}+\mathbf{A})^{-1}\tilde{\eta}G\mathbf{B} \tag{101}$$

$$= \begin{pmatrix}\frac{\eta(1-\beta)G}{2(1+\beta)-\mu\eta(1-\beta)} & -\frac{\eta(1-\beta)G}{2(1+\beta)-\mu\eta(1-\beta)}\end{pmatrix}^\top \tag{102}$$

This yields that:

$$\mathbf{y}_{ZSR}[t] = \mathbf{C}\mathbf{z}[t] = (1 \;\; 0)\,\mathbf{z}[t] \tag{103}$$

$$= (-1)^t \frac{\eta(1-\beta)G}{2(1+\beta) - \mu\eta(1-\beta)} \tag{104}$$

$$\Rightarrow \; \left|\mathbf{y}_{ZSR}[t]\right| = \frac{\eta(1-\beta)G}{2(1+\beta) - \mu\eta(1-\beta)} \tag{105}$$

Lower and Upper bounds. Combining the previous results, we have that $\mathbf{y}[t] = \mathbf{y}_{ZIR}[t] + \mathbf{y}_{ZSR}[t]$. The first term in the r.h.s. starts at θ^0 and under condition in Eq. (97) is converging exponentially to zero. The second term is periodic and the amplitude of the limit cycle increases monotonically with the learning rate η (see Eq. (105)). Choosing a small enough value of η which satisfies the condition (97), *e.g.*

$$\left.\begin{aligned} \eta &> \frac{c_1}{\mu T}\left(\frac{1+\beta}{1-\beta}\right) && \text{(106a)} \\ \eta &< \frac{c_2}{\mu T}\left(\frac{1+\beta}{1-\beta}\right) && \text{(106b)} \end{aligned}\right\} \quad \text{with } T > 1,\; 0 < c_1 < c_2 \leq 2$$

we have that:

$$|\theta^\infty| = \lim_{t\to\infty} |\theta^t| = \lim_{t\to\infty} \big|\underbrace{\mathbf{y}_{ZIR}[t]}_{\text{vanishing}} + \underbrace{\mathbf{y}_{ZSR}[t]}_{\text{periodic}}\big| = \left|\mathbf{y}_{ZSR}[t]\right| \tag{107}$$

$$\theta^\infty \overset{\text{inject (106a) in (105)}}{\geq} \frac{c_1(1+\beta)}{\mu T(1-\beta)} \frac{(1-\beta)G}{2(1+\beta) - \frac{c_1}{T}(1+\beta)} \tag{108}$$

$$= \frac{c_1 G}{\mu(2T - c_1)} \geq \Omega\left(\frac{G}{\mu T}\right) \tag{109}$$

$$\theta^\infty \overset{\text{inject (106b) in (105)}}{\leq} \frac{c_2(1+\beta)}{\mu T(1-\beta)} \frac{(1-\beta)G}{2(1+\beta) - \frac{c_2}{T}(1+\beta)} \tag{110}$$

$$= \frac{c_2 G}{\mu(2T - c_2)} \leq \mathcal{O}\left(\frac{G}{\mu T}\right) \tag{111}$$

We finish the proof by noting that $f(\theta) = \frac{\mu}{2}\theta^2$, with minimum $f(\theta^*) = 0$ at $\theta^* = 0$.

Proof of Thm. 6 (Lower Bound under Decreasing Step-size)

To study the original system from eq. (90), we first split matrix $A[t]$ in two terms:

$$\mathbf{A}[t] = \begin{pmatrix} 1+\beta-\mu\tilde{\eta}_t & -\beta \\ 1 & 0 \end{pmatrix} = \underbrace{\begin{pmatrix} 1+\beta & -\beta \\ 1 & 0 \end{pmatrix}}_{:=\mathbf{A}^\infty} + \underbrace{\begin{pmatrix} -\frac{\mu\tilde{\eta}}{t^\alpha} & 0 \\ 0 & 0 \end{pmatrix}}_{:=\mathbf{E}[t]} \tag{112}$$

With this notation, the system takes the following form:

$$\mathbf{z}[t] = (\mathbf{A}^\infty + \mathbf{E}[t])\,\mathbf{z}[t-1] + \mathbf{B}\mathbf{u}[t] \tag{113}$$

Since the system is time-variant, we cannot directly use the eigenvalues of $\mathbf{A}^\infty$ to analyze its stability and we will need to look at the evolution of the state. To this end, we first transform the system by diagonalizing the part corresponding to $\mathbf{A}^\infty$. We have that

$$\mathbf{A}^\infty = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1} \qquad \mathbf{P} = \begin{pmatrix} 1 & \beta \\ 1 & 1 \end{pmatrix} \qquad \mathbf{\Lambda} = \begin{pmatrix} 1 & 0 \\ 0 & \beta \end{pmatrix} \qquad \mathbf{P}^{-1} = \frac{1}{\beta - 1}\begin{pmatrix} -1 & \beta \\ 1 & -1 \end{pmatrix} \tag{114}$$

and we transform the system (113) as follows:

$$\begin{aligned}
\bar{\mathbf{z}}[t] &= \mathbf{P}^{-1}\mathbf{z}[t] && (115)\\
&= \mathbf{P}^{-1}(\mathbf{A}^\infty + \mathbf{E}[t])\mathbf{P}\bar{\mathbf{z}}[t-1] + \mathbf{P}^{-1}\mathbf{B}\mathbf{u}[t] && (116)\\
&= \big(\underbrace{\mathbf{P}^{-1}\mathbf{A}^\infty\mathbf{P}}_{\mathbf{\Lambda}} + \underbrace{\mathbf{P}^{-1}\mathbf{E}[t]\mathbf{P}}_{:=\mathbf{H}[t]}\big)\bar{\mathbf{z}}[t-1] + \underbrace{\mathbf{P}^{-1}\mathbf{B}}_{:=\mathbf{W}}\mathbf{u}[t] && (117)\\
&= (\mathbf{\Lambda} + \mathbf{H}[t])\bar{\mathbf{z}}[t-1] + \mathbf{W}\mathbf{u}[t] && (118)
\end{aligned}$$

with

$$\begin{aligned}
\mathbf{H}[t] &= -\frac{\mu\tilde{\eta}}{(\beta-1)t^\alpha}\begin{pmatrix} -1 & \beta \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & \beta \\ 1 & 1 \end{pmatrix} && (119)\\
&= -\frac{\mu\tilde{\eta}}{(\beta-1)t^\alpha}\begin{pmatrix} -1 & -\beta \\ 1 & \beta \end{pmatrix} && (120)\\
\mathbf{W} &= \frac{1}{1-\beta}\begin{pmatrix} 1 \\ -1 \end{pmatrix} && (121)
\end{aligned}$$

This leads to:

$$\begin{aligned}
\bar{\mathbf{z}}[t] &= \left[\begin{pmatrix} 1 & 0 \\ 0 & \beta \end{pmatrix} - \frac{\mu\tilde{\eta}}{(1-\beta)t^\alpha}\begin{pmatrix} 1 & \beta \\ -1 & -\beta \end{pmatrix}\right]\bar{\mathbf{z}}[t-1] + \frac{1}{1-\beta}\begin{pmatrix} 1 \\ -1 \end{pmatrix}\mathbf{u}[t] \\
&= \left[\begin{pmatrix} 1 & 0 \\ 0 & \beta \end{pmatrix} - \frac{\mu\eta}{t^\alpha}\begin{pmatrix} 1 & \beta \\ -1 & -\beta \end{pmatrix}\right]\bar{\mathbf{z}}[t-1] + \frac{\eta}{t^\alpha}\begin{pmatrix} 1 \\ -1 \end{pmatrix}G(-1)^t
\end{aligned} \tag{122}$$

where we used the definitions $\tilde{\eta} = (1-\beta)\eta$ and $\mathbf{u}[t] = (-1)^t\tilde{\eta}_t G$ from Lem. 4. We proceed by explicitly writing the transformed-state equation component-wise, i.e.,

$$\begin{cases}
\bar{z}_1[t] = \left(1 - \frac{\mu\eta}{t^\alpha}\right)\bar{z}_1[t-1] - \underbrace{\frac{\mu\eta\beta}{t^\alpha}\bar{z}_2[t-1]}_{:=r_1[t]} + \underbrace{\frac{\eta}{t^\alpha}G(-1)^t}_{:=r_3[t]} & (123a)\\
\bar{z}_2[t] = \left(\beta + \frac{\mu\eta\beta}{t^\alpha}\right)\bar{z}_2[t-1] + \underbrace{\frac{\mu\eta}{t^\alpha}\bar{z}_1[t-1]}_{:=r_2[t]} - \underbrace{\frac{\eta}{t^\alpha}G(-1)^t}_{:=r_3[t]} & (123b)
\end{cases}$$

Now, we unroll these expressions back to the time $t = 0$. Specifically, for $\bar{z}_1[t]$ we have:

$$\bar{z}_1[t] = \left(1 - \frac{\mu\eta}{t^\alpha}\right)\bar{z}_1[t-1] - r_1[t] + r_3[t] \tag{124}$$

$$= \left(1 - \frac{\mu\eta}{t^\alpha}\right)\left[\left(1 - \frac{\mu\eta}{(t-1)^\alpha}\right)\bar{z}_1[t-2] - r_1[t-1] + r_3[t-1]\right] - r_1[t] + r_3[t] \tag{125}$$

$$= \left(1 - \frac{\mu\eta}{t^\alpha}\right)\left(1 - \frac{\mu\eta}{(t-1)^\alpha}\right)\bar{z}_1[t-2] - \left(1 - \frac{\mu\eta}{t^\alpha}\right)r_1[t-1] - r_1[t] \tag{126}$$

$$+ \left(1 - \frac{\mu\eta}{t^\alpha}\right)r_3[t-1] + r_3[t] \tag{127}$$

$$\vdots \tag{128}$$

$$= \prod_{k=2}^{t}\left(1 - \frac{\mu\eta}{k^\alpha}\right)\bar{z}_1[1] + \sum_{s=2}^{t}\prod_{k=s+1}^{t}\left(1 - \frac{\mu\eta}{k^\alpha}\right)\left(r_3[s] - r_1[s]\right) \tag{129}$$

Using similar steps for $\bar{z}_2[t]$ (omitted here for brevity), and defining the shorthand expressions

$$\Psi_1(t, s, \alpha) := \prod_{k=s+1}^{t}\left(1 - \frac{\mu\eta}{k^\alpha}\right) \tag{130}$$

$$\Psi_2(t, s, \alpha) := \prod_{k=s+1}^{t}\left(\beta + \frac{\mu\eta\beta}{k^\alpha}\right) \tag{131}$$

we finally rewrite the original system as

$$\begin{cases} \bar{z}_1[t] = \Psi_1(t, 1, \alpha)\bar{z}_1[1] - \sum_{s=2}^{t}\Psi_1(t, s, \alpha)r_1[s] + \sum_{s=2}^{t}\Psi_1(t, s, \alpha)r_3[s] & \text{(132a)} \\ \bar{z}_2[t] = \Psi_2(t, 1, \alpha)\bar{z}_2[1] + \sum_{s=2}^{t}\Psi_2(t, s, \alpha)r_2[s] - \sum_{s=2}^{t}\Psi_2(t, s, \alpha)r_3[s] & \text{(132b)} \end{cases}$$

Since $\bar{z}_1[t], \bar{z}_2[t]$ are coupled in the system in Eq. (132), in the following we use a technique based on a self-consistent ansatz. That is, we assume an asymptotic form for $\bar{z}_1[t]$ and then verify that the resulting solution for $\bar{z}_2[t]$ leads to a conclusion consistent with the hypothesis. Since the behavior of the system substantially changes when $\alpha > 1$ and $\alpha > 1$, we separately analyze the three cases.

Convergence for $0 < \alpha < 1$. Starting from $\bar{z}_2[t]$, we analyze it assuming $\bar{z}_1[t] \sim c_1(-1)^t/t^\epsilon$, for some arbitrarily small $\epsilon > 0$ and some constant $c_1 > 0$. Under this assumption, from Eq. (132b) we

have that:

$$\lim_{t\to\infty} \bar{z}_2[t] =$$

$$\lim_{t\to\infty} \left[\Psi_2(t,1,\alpha)\bar{z}_2[1] + \mu\eta \sum_{s=2}^{t} \Psi_2(t,s,\alpha) \frac{1}{s^\alpha} \bar{z}_1[s-1] - \eta G \sum_{s=2}^{t} \Psi_2(t,s,\alpha) \frac{(-1)^s}{s^\alpha} \right] \tag{133}$$

$$\overset{11}{=} \lim_{t\to\infty} \left[\beta^t \bar{z}_2[1] - \mu\eta c_1 \sum_{s=2}^{t} \Psi_2(t,s,\alpha) \frac{(-1)^s}{s^{\alpha+\epsilon}} - \eta G \sum_{s=2}^{t} \Psi_2(t,s,\alpha) \frac{(-1)^s}{s^\alpha} \right] \tag{134}$$

$$\overset{14}{=} \lim_{t\to\infty} \left[\beta^t \bar{z}_2[1] - \mu\eta c_1 \frac{(-1)^t}{t^{\alpha+\epsilon}} - \eta G \frac{(-1)^t}{t^\alpha} \right] \tag{135}$$

$$= \lim_{t\to\infty} \eta G \frac{(-1)^{t+1}}{t^\alpha} \tag{136}$$

Where in the second passage we substituted the hypothesis for $\bar{z}_1[t]$ and in third passage we used Lem. 14 twice, with $n = \alpha + \epsilon$ for the second term and $n = \alpha$ for the third term. In the last passage we considered that, since $\epsilon > 0$, the third term is asymptotically slower than both the first and the second. Using the results obtained for $\bar{z}_2[t]$, proceeding from Eq. (132a) we have that:

$$\lim_{t\to\infty} \bar{z}_1[t] =$$

$$\lim_{t\to\infty} \left[\Psi_1(t,1,\alpha)\bar{z}_1[1] - \mu\eta\beta \sum_{s=2}^{t} \Psi_1(t,s,\alpha) \frac{1}{s^\alpha} \bar{z}_2[s-1] + \eta G \sum_{s=2}^{t} \Psi_1(t,s,\alpha) \frac{(-1)^s}{s^\alpha} \right] \tag{137}$$

$$\overset{9}{=} \lim_{t\to\infty} \left[\exp\left(-t^{1-\alpha}\right) \bar{z}_1[1] - \mu\eta^2\beta G \sum_{s=2}^{t} \Psi_1(t,s,\alpha) \frac{(-1)^s}{s^{2\alpha}} + \eta G \sum_{s=2}^{t} \Psi_1(t,s,\alpha) \frac{(-1)^s}{s^\alpha} \right] \tag{138}$$

$$\overset{15}{=} \lim_{t\to\infty} \left[\exp\left(-t^{1-\alpha}\right) \bar{z}_1[1] - \mu\eta^2\beta G \frac{(-1)^t}{t^{2\alpha}} + \eta G \frac{(-1)^t}{t^\alpha} \right] \tag{139}$$

$$= \lim_{t\to\infty} \eta G \frac{(-1)^t}{t^\alpha} \tag{140}$$

Where in the second passage we substituted the result for $\bar{z}_2[t]$ and in third passage we used Lem. 15 twice, with $n = 2\alpha$ for the second term and $n = \alpha$ for the third term. So, for $0 < \alpha < 1$, assuming $\bar{z}_1[t] \sim c_1(-1)^t/t^\epsilon$ leads to the conclusion that $\bar{z}_1[t] \to \eta G(-1)^t/t^\alpha$, so the assumption is valid for $\epsilon = \alpha$ and $c_1 = \eta G$, and any substitution with $\epsilon \in (0, \alpha)$ is valid.

Convergence for $\alpha = 1$. Similarly as before, starting from the assumption $\bar{z}_1[t] \sim (c_1 - c_2(-1)^t)/t^\epsilon$, from Eq. (133) we have that:

$$\lim_{t\to\infty} \bar{z}_2[t] =$$

$$\overset{11}{=} \lim_{t\to\infty} \left[\beta^t \bar{z}_2[1] + \mu\eta \sum_{s=2}^{t} \Psi_2(t,s,1) \frac{c_1 + c_2(-1)^s}{s^{1+\epsilon}} - \eta G \sum_{s=2}^{t} \Psi_2(t,s,1) \frac{(-1)^s}{s} \right] \tag{141}$$

$$\overset{14}{=} \lim_{t\to\infty} \left[\beta^t \bar{z}_2[1] + \mu\eta c_1 \frac{1}{t^{1+\epsilon}} + \mu\eta c_2 \frac{(-1)^t}{t^{1+\epsilon}} - \eta G \frac{(-1)^t}{t} \right] \tag{142}$$

$$= \lim_{t\to\infty} \eta G \frac{(-1)^{t+1}}{t} \tag{143}$$

Using the results obtained for $\bar{z}_2[t]$, proceeding from Eq. (137) we have that:

$$\lim_{t\to\infty} \bar{z}_1[t] =$$

$$\overset{9}{=} \lim_{t\to\infty} \left[\frac{1}{t^{\mu\eta}} \bar{z}_1[1] - \mu\eta^2\beta G \sum_{s=2}^{t} \Psi_1(t,s,1) \frac{(-1)^s}{s^2} + \eta G \sum_{s=2}^{t} \Psi_1(t,s,1) \frac{(-1)^s}{s} \right] \tag{144}$$

$$\overset{15}{=} \lim_{t\to\infty} \left[\frac{1}{t^{\mu\eta}} \bar{z}_1[1] - \mu\eta^2\beta G \frac{(-1)^t}{t^{\mu\eta}} + \eta G \frac{(-1)^t}{t} \right] \tag{145}$$

$$= \lim_{t\to\infty} \begin{cases} \frac{1}{t^{\mu\eta}} \bar{z}_1[1] - \mu\eta^2\beta G \frac{(-1)^t}{t^{\mu\eta}} + \eta G \frac{(-1)^t}{t} & \text{if } \eta \in (0, 1/\mu) \\ \frac{1}{t^{\mu\eta}} \bar{z}_1[1] - \mu\eta^2\beta G \frac{(-1)^t}{t^{\mu\eta}} + \eta G \frac{(-1)^t}{t^{\mu\eta}} & \text{if } \eta \in [1/\mu, 2/\mu) \end{cases} \tag{146}$$

$$= \lim_{t\to\infty} \frac{c_1 - c_2(-1)^t}{t^{\mu\eta}} \tag{147}$$

In particular:

$$c_1 = \bar{z}_1[1], \quad c_2 = \begin{cases} \mu\eta^2\beta G & \text{if } \eta \in (0, 1/\mu) \\ \mu\eta^2\beta G - \eta G & \text{if } \eta \in [1/\mu, 2/\mu) \end{cases} \tag{148}$$

Where in the second passage we used Lem. 15 twice, with $n = 2$ for the second term and $n = 1$ for the third term, and considered the constraint $\mu\eta < 2$. In conclusion, for $\alpha = 1$, assuming $\bar{z}_1[t] \sim (c_1 - c_2(-1)^t)/t^\epsilon$ leads to the conclusion that $\bar{z}_1[t] \to (c_1 - c_2(-1)^t)/t^{\mu\eta}$, so the assumption is valid for $\epsilon = \mu\eta$ and c_1, c_2 as above, and any substitution with $\epsilon \in (0, \mu\eta)$ is valid.

Let us notice that, while it is possible to make $c_2 = 0$ (*i.e.* independent on G) by choosing $\beta = (\mu\eta)^{-1}$ and $\eta > 1/\mu$ (otherwise resulting in the incompatible requirement $\beta = 1$), this does not result in overcoming the dependence on G for the original state $z_1[t]$. In fact, since $z_1[t] = \bar{z}_1[t] + \beta\bar{z}_2[t]$ (Eq. (156)), for $\eta > 1/\mu$ $\bar{z}_2[t]$ in Eq. (143) dominates the rate.

Convergence for $\alpha > 1$. Similarly as before, starting from the assumption $\bar{z}_1[t] \sim c_1$, from Eq. (133) we have that:

$$\lim_{t\to\infty} \bar{z}_2[t] =$$

$$\overset{11}{=} \lim_{t\to\infty} \left[\beta^t \bar{z}_2[1] + \mu\eta c_1 \sum_{s=2}^{t} \Psi_2(t,s,\alpha) \frac{1}{s^\alpha} - \eta G \sum_{s=2}^{t} \Psi_2(t,s,\alpha) \frac{(-1)^s}{s^\alpha} \right] \tag{149}$$

$$\overset{14}{=} \lim_{t\to\infty} \left[\beta^t \bar{z}_2[1] + \mu\eta c_1 \frac{1}{t^\alpha} - \eta G \frac{(-1)^t}{t^\alpha} \right] \tag{150}$$

$$= \lim_{t\to\infty} \frac{\mu\eta c_1 + \eta G(-1)^{t+1}}{t^\alpha} \tag{151}$$

Using the results obtained for $\bar{z}_2[t]$, proceeding from Eq. (137) we have that:

$$\lim_{t\to\infty} \bar{z}_1[t] =$$

$$\overset{9}{=} \lim_{t\to\infty} \left[c\bar{z}_1[1] - \mu\eta\beta \sum_{s=2}^{t} \Psi_1(t,s,\alpha) \frac{\mu\eta c_1 + \eta G(-1)^s}{s^{2\alpha}} + \eta G \sum_{s=2}^{t} \Psi_1(t,s,\alpha) \frac{(-1)^s}{s^{\alpha}} \right] \tag{152}$$

$$\overset{12+15}{=} \lim_{t\to\infty} \left[c\bar{z}_1[1] - \mu\eta\beta c_1 g(2\alpha) - \mu\eta\beta f(2\alpha)\eta G + \eta G f(\alpha) \right] \tag{153}$$

where c is a positive constant as in Corollary 11 and $g(\alpha), f(\alpha)$ are functions in α, constant in t, determining the proper value at convergence of $\bar{z}_1[t]$ (as bounded in Lems. 12 and 15). Assuming initialization at optimum (*i.e.* $\bar{z}_1[1] = 0$), we solve for c_1:

$$\lim_{t\to\infty} \bar{z}_1[t] = c_1 = \eta G \frac{f(\alpha) - \mu\eta\beta f(2\alpha)}{1 + \mu\eta\beta g(2\alpha)} \tag{154}$$

$$\overset{\eta\sim\mathcal{O}(1/\mu)}{=} \Theta\left(\frac{G}{\mu}\right) \tag{155}$$

Convergence of the original system $\mathbf{z}[t]$. Recalling that $\mathbf{z}[t] = \mathbf{P}\bar{\mathbf{z}}[t]$, we have that:

$$\mathbf{z}[t] = \begin{pmatrix} 1 & \beta \\ 1 & 1 \end{pmatrix} \bar{\mathbf{z}}[t] = \begin{pmatrix} \bar{z}_1[t] + \beta\bar{z}_2[t] \\ \bar{z}_1[t] + \bar{z}_2[t] \end{pmatrix} \tag{156}$$

So, from Eq. (136) and (140),Eq. (143) and (147) and from Eq. (151) and (155), we have that:

$$\lim_{t\to\infty} |z_1[t]| = \begin{cases} \Theta\left(\dfrac{G}{\mu t^{\alpha}}\right) & \text{if } 0 < \alpha < 1 \\ \Theta\left(\dfrac{G}{\mu t^{\min(\mu\eta,1)}}\right) & \text{if } \alpha = 1 \\ \Theta\left(\dfrac{G}{\mu}\right) & \text{if } \alpha > 1 \end{cases} \tag{157}$$

We finish the proof by noting that $f(\theta) = \frac{\mu}{2}\theta^2$, with minimum $f(\theta^*) = 0$ at $\theta^* = 0$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All the claims are reflected into the contents of the paper

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, we discuss in the paper that our claims only apply to classical momentum, and mention explicitly algorithm our paper does not apply to (Sec. A.2).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution

is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All the necessary assumptions are stated in the corresponding section, and full proof are provided in the appendix

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental details are provided in Appendix B

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code is not provided since the OpenReview form has not enabled option for supplementary material. Provided details are sufficient to reproduce the experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides sufficient details for reproducibility

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Where relevant, experiments are repeated multiple times and standard deviation reported (i.e. Figure 1).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Yes

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA] .

    Justification: This paper outlines limits of optimization, there is societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: Safeguards do not apply to this paper.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA] .

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] .

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA] .

    Justification: This research does not involve LLMs

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.